

# VU Research Portal

## Of Zeros and Ones

Rauschenberger, A.

2021

### **document version**

Publisher's PDF, also known as Version of record

[Link to publication in VU Research Portal](#)

### **citation for published version (APA)**

Rauschenberger, A. (2021). *Of Zeros and Ones: Association and Prediction in Genomics*. [PhD-Thesis - Research and graduation internal, Vrije Universiteit Amsterdam].

### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

### **E-mail address:**

[vuresearchportal.ub@vu.nl](mailto:vuresearchportal.ub@vu.nl)

# Contents

<b>1</b>	<b>Introduction</b>	<b>11</b>
1.1	Molecular biology . . . . .	12
1.2	High-dimensional statistics . . . . .	14
1.3	Experimental data . . . . .	17
1.4	Overview . . . . .	18
<b>2</b>	<b>Negative binomial global test</b>	<b>23</b>
2.1	Background . . . . .	26
2.2	Methods . . . . .	27
2.2.1	The random-effects model . . . . .	27
2.2.2	The testing procedure . . . . .	28
2.2.3	Relation to the Poisson model . . . . .	30
2.2.4	Individual contributions . . . . .	30
2.2.5	Method of control variables . . . . .	31
2.2.6	Multiple molecular profiles . . . . .	32
2.3	Results . . . . .	33
2.3.1	Simulation study . . . . .	33
2.3.2	Application: HapMap . . . . .	35
2.3.3	Stratified permutation test . . . . .	35
2.3.4	Presence of overdispersion . . . . .	36
2.3.5	Individual contributions . . . . .	36
2.3.6	Application: TCGA . . . . .	37
2.3.7	Robustness to multicollinearity . . . . .	40
2.3.8	Method of control variables . . . . .	40

2.3.9	Multiple molecular profiles . . . . .	41
2.4	Discussion . . . . .	42
2.5	Appendix . . . . .	46
2.5.1	Derivation of the test statistic . . . . .	46
2.5.2	Cancer dataset . . . . .	48
2.5.3	Additional figures and tables . . . . .	49
<b>3</b>	<b>Semi-supervised mixture test</b>	<b>61</b>
3.1	Background . . . . .	64
3.2	Methods . . . . .	66
3.2.1	Semi-supervised mixture model . . . . .	66
3.2.2	Main and interactive effects in genetics . . . . .	70
3.3	Simulation . . . . .	73
3.3.1	Data generating process . . . . .	73
3.3.2	Statistical power . . . . .	74
3.3.3	False positive rate . . . . .	78
3.4	Application . . . . .	79
3.4.1	GWAS . . . . .	79
3.4.2	eQTLs . . . . .	81
3.5	Discussion . . . . .	84
3.6	Appendix . . . . .	87
3.6.1	Framework . . . . .	87
3.6.2	Expectation-maximisation algorithm . . . . .	88
3.6.3	Gaussian distribution . . . . .	91
3.6.4	Negative binomial distribution . . . . .	92
3.6.5	Additional figures . . . . .	93
<b>4</b>	<b>Multinomial global test</b>	<b>105</b>
4.1	Background . . . . .	108
4.2	Methods . . . . .	110
4.2.1	The model . . . . .	110
4.2.2	Test statistic for splice-changing events . . . . .	113
4.2.3	Relation with other works . . . . .	115
4.2.4	Experimental data used in examples . . . . .	115
4.3	Results . . . . .	117

4.3.1	Simulation study . . . . .	117
4.3.2	Example 1: GEUVADIS . . . . .	119
4.3.3	Example 2: GEUVADIS and LLS . . . . .	123
4.4	Discussion . . . . .	125
4.5	Appendix . . . . .	130
4.5.1	Derivation of the score test statistic . . . . .	130
4.5.2	Related test statistics . . . . .	137
4.5.3	Simulation studies . . . . .	141
4.5.4	Pre-processing of example data . . . . .	145
4.5.5	Sequential testing . . . . .	145
4.5.6	Additional figures and tables . . . . .	147
<b>5</b>	<b>Paired lasso</b>	<b>163</b>
5.1	Background . . . . .	166
5.2	Methods . . . . .	167
5.2.1	Setting . . . . .	167
5.2.2	Paired lasso . . . . .	169
5.2.3	Initial estimators . . . . .	171
5.3	Results . . . . .	173
5.3.1	Classification problems . . . . .	173
5.3.2	Paired covariates . . . . .	175
5.3.3	Predictive performance . . . . .	176
5.3.4	Weighting schemes . . . . .	178
5.4	Discussion . . . . .	185
5.5	Appendix . . . . .	189
5.5.1	Background . . . . .	189
5.5.2	Methods . . . . .	190
5.5.3	Results . . . . .	191
5.5.4	Discussion . . . . .	195
<b>6</b>	<b>Stacked elastic net</b>	<b>203</b>
6.1	Background . . . . .	206
6.2	Methods . . . . .	207
6.2.1	Base learners . . . . .	207
6.2.2	Meta learner . . . . .	208

6.2.3	Combination . . . . .	209
6.2.4	Extension . . . . .	210
6.3	Simulation . . . . .	212
6.3.1	Prediction accuracy . . . . .	212
6.3.2	Estimation accuracy . . . . .	214
6.4	Application . . . . .	217
6.4.1	Benchmark data sets . . . . .	217
6.4.2	Normal/tumour classification . . . . .	218
6.5	Discussion . . . . .	219
<b>7</b>	<b>Discussion</b>	<b>225</b>
7.1	Common themes . . . . .	226
7.2	Performance evaluation . . . . .	226
7.3	Approximate permutation testing . . . . .	227
7.4	Distribution of $p$ -values . . . . .	228
7.5	Variable selection . . . . .	228
7.6	Association and prediction . . . . .	229
7.7	Future research . . . . .	230
	<b>Summary</b>	<b>231</b>
	<b>Acknowledgements</b>	<b>234</b>
	<b>Abbreviations</b>	<b>235</b>
	<b>Bibliography</b>	<b>249</b>